L'ombre d'un doute ? Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles

Lucie Scarbel, Denis Beautemps, Jean-Luc Schwartz, S. Schmerber, Marc Sato

L'ombre d'un doute?
Interactions perceptivo-motrices lors de tâches de close-shadowing auditive et audio-visuelles

Lucie Scarbel[1], Denis Beautemps [1], Jean-Luc Schwartz [1],
Sébastien Schmerber[2], Marc Sato [1]
(1) GIPSA-LAB, Département Parole & Cognition ; CNRS UMR 5216 - Université de Grenoble-Alpes ;
(2) Service ORL du CHU Grenoble
lucie.scarbel@gipsa-lab.grenoble-inp.fr

RESUME

Un argument classique en faveur des théories motrices de la perception de la parole provient du paradigme de « close-shadowing » (répétition rapide). Le fait que cette tâche de close-shadowing entraîne des réponses orales bien plus rapides qu'en réponses manuelles suggère en effet un codage des représentations perceptives dans un format moteur, compatible avec une réponse orale. Un autre argument est apporté par les interactions audio-visuelles lors de la perception de parole, souvent interprétées en référence à un couplage fonctionnel entre audition, vision et motricité. Dans cette étude, nous avons combiné ces deux paradigmes de manière à tester si la modalité visuelle pouvait induire des réponses motrices plus rapides lors d'une tâche de close-shadowing. Pour ce faire, différentes tâches de catégorisation orale et manuelle de stimuli de parole présentés auditivement ou audio-visuellement, en présence ou non d'un bruit blanc, ont été réalisées. De manière générale, les réponses orales ont été plus rapides que les réponses manuelles, mais aussi moins précises, notamment dans le bruit, ce qui suggère que la représentation motrice induite par la stimulation pourrait être peu précise dans un premier niveau de traitement. En présence d'un bruit acoustique, la modalité audiovisuelle s'est avérée à la fois plus rapide et plus précise que la modalité auditive. Aucune interaction entre le mode de réponse et la modalité de présentation des stimuli n'a cependant été observée. Nous interprétons l'ensemble de ces résultats dans un cadre théorique proposant l'existence de boucles perceptivo-motrices, dans lesquelles les entrées auditives et visuelles seraient intégrées et reliées à la génération interne de représentations motrices préalablement au processus final de décision.

ABSTRACT

The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close shadowing

One classical argument in favor of a functional role of the motor system in speech perception comes from the close shadowing task in which a subject has to identify and to repeat as quickly as possible an auditory speech stimulus. The fact that close shadowing can occur very rapidly and much faster than manual identification of the speech target is taken to suggest that perceptually-induced speech representations are already shaped in a motor-compatible format. Another argument is provided by audiovisual interactions often interpreted as referring to a multisensory-motor framework. In this study, we attempted to

combine these two paradigms by testing whether the visual modality could speed motor response in a close-shadowing task. To this aim, both oral and manual responses were evaluated during the perception of auditory and audio-visual speech stimuli, clear or embedded in white noise. Overall, oral responses were much faster than manual ones, but it also appeared that they were less accurate in noise, which suggests that motor representations evoked by the speech input could be rough at a first processing stage. In the presence of acoustic noise, the audiovisual modality led to both faster and more accurate responses than the auditory modality. No interaction was however observed between modality and response. Altogether, these results are interpreted within a two-stage sensory-motor framework, in which the auditory and visual streams are integrated together and with internally generated motor representations before a final decision may be available.

MOTS-CLES : perception de la parole, production de la parole, perception de la parole audio-visuelle, close-shadowing, interaction sensorimotrice

KEYWORDS: speech perception, speech production, audio visual speech perception, close-shadowing, sensory motor interaction

1 Introduction

Un débat classique dans le domaine de la perception de la parole concerne l'implication du système moteur et du lien fonctionnel entre représentations auditives et motrices. Les théories auditives de la perception de parole réfutent l'implication du système moteur et proposent l'existence de processus auditifs de décodage phonétique à partir du signal acoustique (Diehl et al., 2004). A contrario, Liberman et al. (1985) et Fowler (1986) supposent que pour percevoir la parole, nous utilisons des représentations procédurales motrices basées sur notre expérience de locuteur. Enfin, les théories perceptivo-motrices postulent que les représentations motrices soient utilisées en lien avec les représentations auditives dans le traitement et décodage des informations phonétiques (Skipper et al., 2007, Schwartz et al. 2012).

Dans une récente revue sur les théories motrices, Galantucci et al. (2006) rappellent les principaux arguments expérimentaux et en mentionnent notamment deux qui fournissent le cœur du présent travail. Un premier argument en faveur d'un couplage perceptivo-moteur en parole provient du paradigme de close-shadowing. Ce paradigme (Porter et al., 1980 ; Fowler et al., 2003) consiste à répéter le plus rapidement possible des stimuli de parole. L'analyse des réponses orales permet de mesurer la rapidité et le taux de répétitions correctes. Galantucci (2006) compare ces résultats avec les temps de réaction obtenus par Luce (1986) dans une tâche de réponse manuelle dans laquelle les sujets utilisaient une touche. Les résultats montrent une augmentation des temps de réaction dans le cas de la réponse manuelle. Cette différence ne pouvant être expliquée par la difficulté du choix des touches dans la tâche de décision manuelle, Galantucci et al. l'interprètent par le fait que si percevoir la parole c'est percevoir des gestes, alors la perception des gestes préparerait la réponse orale et la rendrait ainsi plus rapide.

Le second paradigme est celui de la multisensorialité. De nombreuses études ont montré que l'entrée visuelle améliore la compréhension de parole, tant pour les sujets malentendants que normo-entendants. Sumby & Pollack (1954) ont été parmi les premiers à démontrer l'apport de la modalité visuelle pour percevoir et comprendre la parole dans des conditions

bruitées. L'apport de la modalité visuelle a été également démontré dans des études de répétition (shadowing) où la tâche consistait à répéter oralement les stimuli, et à mettre ainsi en évidence une meilleure compréhension par des sujets de matériaux linguistiques complexes ou produits dans une langue étrangère ou avec accent étranger (Reisberg, 1987 ; Davis and Kim, 2001). Cependant, ces tâches de répétition se sont faites sans pression de temps (shadowing et non close-shadowing). Or la tâche de répétition rapide fournit une fenêtre riche sur la dynamique temporelle du processus de décision. A l'inverse les expériences de close-shadowing n'ont jamais incorporé la modalité visuelle, se privant d'une connaissance sur le rôle des interactions audiovisuelles en lien avec les relations perceptuo-motrices. La présente étude se propose précisément d'étudier, pour la première fois, quel est l'apport de la modalité visuelle dans une tâche de close-shadowing.

Cette étude est composée de deux expériences, toutes deux focalisées sur une évaluation conjointe de la précision et de la rapidité de réponses orales ou manuelles à des stimuli auditifs ou audio-visuels. Les deux expériences ont été réalisées sur des stimuli de parole non-lexicaux (logatomes), présentés sans bruit dans la première expérience (Expérience A) ou avec bruit acoustique dans la seconde expérience (Expérience B). Les hypothèses sous-jacentes sont que (1) les réponses orales devraient être plus rapides que les réponses manuelles, en accord avec les études précédentes sur le close-shadowing, et que (2) les réponses aux stimuli audio-visuels devraient être plus rapides et plus précises que celles aux stimuli auditifs, au moins dans l'Expérience B impliquant des stimuli bruités. Une question supplémentaire concerne la possibilité d'une interaction entre ces deux effets, qui permettrait d'évaluer si l'effet de la vision est différent entre une modalité de réponse (orale) et l'autre (manuelle).

2 Méthodologie

2.1 Participants

Deux groupes de quinze et quatorze adultes sains, de langue maternelle française, ont participé aux Expériences A et B (Expérience A: 10 femmes; moyenne d'âge: 29 ans, entre 20 et 38 ans - Expérience B: 11 femmes; moyenne d'âge: 24 ans, entre 19-34 ans). Tous les participants ont une vision normale ou corrigée à la normale et ont rapporté n'avoir jamais eu des troubles moteurs, de la parole ou de l'audition.

2.2 Procédure expérimentale

Chaque expérience consistait en deux tâches de catégorisation : une tâche de close-shadowing où les réponses étaient données oralement, en répétant le plus vite possible la séquence présentée, et une tâche de décision manuelle, où les réponses étaient données manuellement, en appuyant le plus vite possible sur la touche appropriée. Les stimuli à catégoriser correspondaient aux séquences /apa/, /ata/ et /aka/ (voir ci-dessous). Les participants étaient informés qu'on allait leur présenter des séquences /apa/, /ata/ ou /aka/, soit de manière auditive soit de manière audio-visuelle. Dans la tâche de close-shadowing, on leur demandait de catégoriser et répéter chaque séquence le plus vite possible. Pour ce faire, ils devaient produire la voyelle initiale /a/ puis répéter immédiatement la syllabe CV perçue (/pa/, /ta/ ou /ka/). Lors de la tâche de décision manuelle, les participants devaient catégoriser chaque énoncé en appuyant le plus vite

possible avec leur main dominante sur une des trois touches correspondant respectivement à /apa/, /ata/ ou /aka/. L'ordre des touches était contrebalancé entre les participants. Pour chaque tâche (avec réponses manuelles ou orales) et chaque modalité (auditive ou audio-visuelle), 16 répétitions de chacune des séquences /apa/, /ata/ et /aka/ étaient présentées de manière randomisée. Les ordres de présentation des tâches et des modalités étaient contrebalancés entre les participants. Les deux expériences ont été réalisées dans une chambre sourde. Les participants étaient assis en face d'un ordinateur à une distance d'approximativement 50 cm. Les stimuli acoustiques étaient présentés à un niveau sonore confortable, celui-ci étant le même pour tous les participants. Le logiciel Presentation (Neurobehavioral Systems, Albany, CA) a été utilisé pour contrôler la présentation des stimuli et pour enregistrer les réponses manuelles. Toutes les productions des participants ont été enregistrées grâce à un microphone AKG 1000S pour les analyses offline, avec un système assurant la synchronisation entre les stimuli présentés et la réponse des participants. Une courte session d'entrainement précédait chaque tâche. La durée totale de chaque expérience était d'environ 30 minutes.

2.3 Stimuli

Les séquences /apa/ /ata/ et /aka/, produites dans une chambre sourde par un homme de langue maternelle française ont été enregistrées audio-visuellement au moyen d'un microphone AKG 1000S (44.1 kHz) et d'une caméra haute qualité au format PAL placée en face du locuteur (images détramées de 572 par 520 pixels, 50 Hz). Le corpus a été enregistré avec pour objectif d'obtenir pour chaque séquence 4 productions distinctes impliquant quatre durées de la voyelle initiale /a/ (0.5s, 1s, 1.5s et 2 s) dans le but de réduire toute prédiction temporelle de la séquence à catégoriser). Les durées des 12 stimuli ainsi sélectionnés ont été égalisées (3 séquences x 4 productions). Les stimuli étaient présentés sans bruit additionnel dans l'Expérience A, un bruit blanc (filtré à -6 dB/oct) a été ajouté à chaque stimulus dans l'Expérience B (rapport signal sur bruit de -3 dB).

2.4 Analyses acoustiques

Afin de calculer les temps de réaction (RTs) et la proportion de réponses correctes dans la tâche à réponses orales, des analyses acoustiques de la production des participants ont été réalisées en utilisant le logiciel Praat (Boersma et Weenink, 2013). Les temps de réaction ont été calculés uniquement pour les réponses correctes : les omissions ou tout autre type d'erreurs (c'est-à-dire le remplacement d'une consonne par une autre ou la production de deux consonnes ou de deux syllabes pour un même stimulus dans la tâche de close-shadowing) ont été exclues. Les temps de réaction pour les réponses orales ont été mesurés entre le début du burst consonantique du stimulus et de la réponse.

2.5 Analyse des données

Dans chaque expérience, la proportion de réponses correctes et la médiane des temps de réaction étaient déterminées individuellement pour chaque participant, chaque tâche et chaque modalité, séparément pour /apa/, /ata/ et /aka/. Deux ANOVA à mesures répétées ont été réalisées sur ces données, avec le groupe (Expérience A avec stimuli non-bruités vs. Expérience B avec stimuli bruités) comme variable inter-sujets et la tâche (close shadowing

vs. décision manuelle), la modalité (audio vs. audio-visuelle), et le type de stimulus (/apa/ vs /ata/ vs /aka/) comme variables intra-sujets.

3 Résultats

Pour toutes les analyses suivantes, le niveau de significativité était fixé à $p = .05$ et corrigé en cas de violation de l'hypothèse de sphéricité. Les analyses post-hoc ont été effectuées en utilisant des tests de Bonferroni.

3.1 Temps de réaction

Comme attendu, l'effet principal du groupe est significatif ($F(1,27)=24,38$; $p<0.001$), avec des temps de réaction pour les stimuli non bruités de l'expérience A plus courts que ceux des stimuli bruités de l'expérience B (351 ms vs 484 ms). Les effets principaux de la tâche ($F(1,27)=151,70$; $p<0.001$) et de la modalité ($F(1,27)=14,79$; $p<0.001$) sont aussi significatifs. Pour la tâche, les réponses orales étaient plus rapides que les réponses manuelles (286ms vs 545ms). Par rapport à la modalité, les temps de réaction étaient plus courts dans la modalité audio-visuelle par rapport à la modalité auditive (405 ms vs. 425 ms). Une interaction significative entre groupe et modalité ($F(1,27)=21,74$; $p<0.001$) montre que l'effet bénéfique de la présentation audio-visuelle est présent avec les stimuli bruités dans l'expérience B (461 ms vs. 507 ms) mais non avec les stimuli non-bruités de l'expérience A (354 ms vs. 349 ms). Par contre, l'interaction entre modalité et réponse n'est pas significative.

Ces effets semblent être dépendants des syllabes perçues. Notamment, une interaction à trois facteurs 'tâche x modalité x syllabe' a été trouvée ($F(2,54)=6,49$; $p<0.005$). Dans la modalité auditive, aucune différence significative des temps de réaction n'a été observée entre les syllabes à la fois pour les réponses orales et pour les réponses manuelles. Par contre, dans la modalité audio-visuelle, les temps de réaction pour les réponses orales étaient plus rapides pour la syllabe /pa/ par rapport aux syllabes /ta/ et /ka/, alors que les temps de réaction pour les réponses manuelles étaient plus rapides pour /pa/ par rapport à /ka/ et pour /ka/ par rapport à /ta/.

Ainsi, globalement, on obtient un patron de résultats attendus : des temps de réponse plus courts en réponse orale, un effet du bruit ralentissant les temps de réponse, une accélération de la réponse en modalité audiovisuelle par rapport à la modalité auditive en présence de bruit. Il n'apparaît cependant pas d'interaction entre modalité et type de réponse.

3.2 Proportion de réponses correctes

L'effet principal du groupe est significatif ($F(1,27)=266,28$; $p<0.001$) avec une proportion de réponses correctes plus élevée pour les stimuli non-bruités de l'expérience A (95%) par rapport aux stimuli bruités de l'expérience B (61%). D'autres effets principaux ont été significatifs, à la fois pour la tâche ($F1(1,27)=69,40$; $p<0.001$) et pour la modalité ($F(1,27)=52,39$; $p<0.001$). Concernant la tâche, une baisse importante des réponses correctes a été observée pour les réponses orales par rapport aux réponses manuelles (73% vs. 85%). Comme l'interaction significative du groupe par la tâche ($F(1,27)=38,67$; $p<0.001$) l'indique, cet effet n'apparait que pour les stimuli bruités de l'expérience B (71%

vs. 50%) alors qu'aucune différence n'a été observée entre les réponses orales et manuelles pour les stimuli non-bruités de l'expérience A (93% vs. 98%). Concernant la modalité, la modalité audio-visuelle apporte plus de réponses correctes que la modalité auditive (82% vs. 75%). Par contre, comme l'indique l'interaction significative 'groupe x modalité' ($F(1,27)=72,36$; $p<0.001$) aucune différence n'apparaît entre les deux modalités avec les stimuli non-bruités de l'expérience A (96% vs. 95%) alors qu'avec les stimuli bruités de l'expérience B, la modalité audio-visuelle apporte plus de réponses correctes (68%) que la modalité auditive (53%).

Là encore, les résultats dépendent de la syllabe présentée. Si les 3 syllabes sont parfaitement identifiées dans l'Expérience A en l'absence de bruit, quelle que soit la tâche et la modalité, en condition bruitée (Expérience B) la syllabe « pa » apparaît la plus saillante à la fois auditivement et visuellement. Si on obtient ici encore une confirmation d'un patron attendu (réponses plus précises en l'absence de bruit et, dans le cas de stimuli bruités, en présence de la modalité visuelle), un résultat fort et inattendu doit être relevé : le fait que la réponse orale dégrade la précision de la réponse en cas de stimuli bruités (Expérience B). Une nouvelle fois, il n'apparaît pas d'interaction entre modalité et type de réponse.

Tableau 1: Moyenne des temps de réaction (en ms.) et des % de réussite

modalité	A	A	A	A	A	A	AV	AV	AV	AV	AV	AV
mode	oral	oral	oral	manuel	manuel	manuel	oral	oral	oral	manuel	manuel	manuel
syllabe	ka	pa	ta	ka	pa	ta	ka	pa	ta	ka	pa	ta
RTs sans bruit	250	208	259	471	442	465	261	197	268	474	416	506
RTs avec bruit	348	335	373	614	666	632	277	296	318	531	653	601
% sans bruit	90%	99%	93%	98%	99%	98%	89%	100%	88%	98%	96%	97%
% avec bruit	56%	36%	40%	67%	42%	79%	89%	44%	38%	96%	58%	85%

4 Discussion

4.1 Effet de la tâche : mode de réponse oral ou manuel

Dans la condition non bruitée (Expérience A), les réponses orales sont plus rapide que les réponses manuelles (240ms vs. 462ms), mais aussi marginalement moins précises (93% vs. 98%, effet non significatif). Les temps de réaction sont en adéquation avec ceux obtenus par Fowler et al. (2003) et par Porter & Castellanos (1980). Ces auteurs interprètent la rapidité de la réponse dans le mode oral en référence avec les théories motrices. Le système orofacial serait ainsi favorisé pour répondre de manière très rapide, puisque le percept serait déjà en adéquation avec le format moteur ; le système manuel, nécessitant une étape transitoire entre la décision et l'action, serait ralenti d'autant. Cependant, les données de la condition bruitée (Expérience B) apportent des précisions importantes et inattendues sur ce raisonnement. En effet, alors que les temps de réaction restent plus courts dans la tâche de close-shadowing (334ms vs. 633ms), la précision dans la tâche de réponse orale diminue considérablement par rapport à la tâche de décision manuelle (50% vs. 71%).

Ces nouvelles données nécessitent de modifier l'interprétation des tenants des théories motrices jusqu'à un certain point. Nous allons proposer une tentative d'explication dans le

cadre du modèle proposé par Skipper et al. (2007) pour intégrer les interactions perceptuo-motrices dans la perception de parole. Ces auteurs proposent un modèle inspiré de « l'analyse par la synthèse » (Stevens & Halle 1967 ; Bever & Poeppel, 2010). Le modèle de Skipper et al. implique une boucle corticale entre les aires auditives et motrices. Après un stade initial de traitement auditif dans le cortex temporal (cortex auditif primaire, secondaire et aires associatives : stade 1), le cortex frontal générerait des hypothèses phonémiques associées avec les buts articulatoires puis des commandes motrices correspondant à cette prédiction initiale (Pars opercularis, cortex pré moteur ventral et cortex moteur primaire : stade 2), afin d'émettre des copies d'efférence qui seraient ensuite renvoyées dans le cortex auditif afin d'être comparées avec l'input auditif (stade 3). Ce modèle peut être utilisé comme une base pour tenter d'interpréter nos propres données. Pour ce faire, nous supposons que les réponses manuelles et orales sont générées à deux stades différents dans cette boucle de traitement. Les réponses orales seraient générées au stade 2, en accord avec les postulats de Porter & Castellanos ou de Fowler et al. Quand l'information provenant du cortex auditif aurait généré des commandes motrices dans le cortex moteur, le système orofacial, pré-activé depuis le début de l'expérience de close-shadowing pour permettre aux participants de répondre le plus vite possible, générerait une réponse orale produite par ces commandes motrices. Les réponses orales étant traitées à un stade précoce, elles seraient donc plus rapides mais sont aussi, selon nos résultats, moins précises, ce qui correspond au modèle de Skipper et al. qui considère qu'il ne s'agirait que d'une première hypothèse de réponse qui devraient être affinées à un stade ultérieur. Au stade 2, par contre, le système manuel ne reçoit pas de stimulations spécifiques permettant de générer une réponse. Par contre, au stade suivant (stade 3), le transfert de l'information auditive au cortex auditif, grâce à la copie d'efférence, fournit, en intégrant cette hypothèse motrice avec l'entrée acoustique, une information plus précise qui peut alors être transférée au système manuel pour réponse. Dans ce raisonnement, les réponses orales et manuelles seraient émises à deux instants différents du processus d'analyse par synthèse. Par conséquent, les temps de réaction pour les réponses manuelles seraient plus lents que ceux des réponses orales, mais les réponses seraient plus précises puisque, contrairement aux réponses orales, dans la tâche de décision manuelle les prédictions auraient été confirmées et ajustées avec le cortex auditif avant la décision finale envoyée aux commandes motrices manuelles pour appui de la touche adéquate. Bien sûr cette explication est probablement trop simple pour tenir compte de tous les aspects de nos données. Néanmoins, l'aspect crucial de nos résultats est qu'une hypothèse de pur processus d'identification motrice, certes compatible avec des temps de réaction plus courts en réponse orale, ne semble pas pouvoir rendre compte de la diminution de la précision de réponse orale pour des stimuli bruités. Ces résultats semblent donc réfuter une version stricte des théories motrices au profit de théories perceptuo-motrices de la perception de parole telles que celle de Skipper et al. (2007).

4.2 Effets de la modalité : auditive vs. audio-visuelle

Les effets de la modalité n'apparaissent dans notre étude que dans l'Expérience B avec les stimuli bruités. Dans la modalité auditive, les temps de réaction sont alors plus longs que dans la modalité audio-visuelle, et les proportions de réponses correctes sont plus faibles. Pris ensemble, ces résultats montrent un bénéfice clair de l'apport de la modalité visuelle à l'input auditif, ce qui est en accord avec toutes les études depuis Sumby et Pollack (1954).

Dans notre étude, l'avantage audio-visuel apparaît essentiellement pour la syllabe /pa/ ce qui est classique et en lien avec la haute visibilité des mouvements des lèvres associées à la bilabiale /p/, et le fort degré de confusion entre les mouvements visuels associés à /t/ ou /k/, généralement considérés comme appartenant à la même classe visémique. Ces effets de la modalité ne sont pas présents dans l'Expérience A avec des stimuli non-bruités, très probablement parce que les temps de réaction dans la modalité auditive sont déjà trop courts et les proportions de réponses correctes trop élevées pour être améliorés par l'entrée visuelle (effet plafond).

Un point intéressant est qu'il n'y a pas d'interaction significative entre la modalité et la tâche c'est-à-dire que la diminution des temps de réaction et l'amélioration des proportions de réponses correctes de la modalité auditive à la modalité audiovisuelle sont similaires dans les tâches orales ou manuelles. Nous allons là encore tenter d'interpréter cette absence d'interaction dans le cadre du modèle proposé par Skipper et collègues. Dans ce modèle, les informations auditives et visuelles, après prétraitement dans les aires auditives et visuelles, convergeraient dans une aire multi-sensorielle dans le cortex temporal postéro-supérieur (stade 1). Ensuite, dans le cas d'un input multi-sensoriel, les premières hypothèses seraient donc plutôt multi-sensorielles plutôt qu'uniquement auditives. A partir de là comme précédemment génération d'une hypothèse phonémique associée avec des buts articulatoires puis des commandes motrices orofaciales (stade 2), et émission d'une copie d'efférence fournissant une prédiction multisensorielle comparée avec l'input (stade 3). Dans notre étude, les interactions audio-visuelles du stade 1 affineraient les procédés sensori-moteurs et produiraient des hypothèses phonémiques plus rapides et plus précises au stade 2, qui est le stade où, dans notre interprétation, les réponses orales seraient générées. Ensuite le même gain en rapidité et en précision serait rétro-propagé vers les aires auditives pour génération des réponses manuelles (stade 3). Il n'y aurait ainsi pas de raison d'attendre des différences de gain visuel entre les tâches orales ou manuelles, le gain étant essentiellement déterminé dès le stade 1 dans le modèle.

5 Conclusion

En résumé, les résultats de la présente étude suggèrent que les réponses manuelles et orales sont générées à deux stades différents dans la chaine de perception de la parole. Dans la théorie du modèle « d'analyse par synthèse », les réponses manuelles serait fournies seulement à la fin de la boucle complète, incluant un processus feedworward de génération de prédictions phonémiques associées avec les buts articulatoires puis des commandes motrices émettant en feedback des hypothèses multi-sensorielles comparées au flux d'évènements multi-sensoriels. Contrairement aux réponses manuelles, les réponses orales seraient produites à un stade plus précoce (fin du processus feedforward) où les commandes motrices sont générées, produisant des réponses plus rapides mais moins précises. L'apport visuel améliorerait la rapidité et la précision pour les phonèmes suffisamment visibles (comme par exemple /p/) et ce lorsque le système auditif est en difficulté (conditions adverses, par exemple en présence de bruit). Bien évidemment, d'autres interprétations ou d'autres théories pourraient être confrontées aux présentes données expérimentales. Mais globalement, l'ensemble des résultats de cette étude semble nécessiter une théorie perceptuo-motrice de la perception de parole dans laquelle les flux auditifs et visuels sont intégrés à des représentations motrices auto-générées, avant d'aboutir à une décision finale.

Remerciements

Les auteurs tiennent à remercier l'ANR Plasmody qui a permis de financer cette étude, ainsi que les participants qui se sont portés volontaires pour les deux expériences.

Bibliographie

Bever, T. G., & Poeppel, D. (2010). Analysis by synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics,* 4.2-.3, 174-200.

Davis, C. & Kim, J. (2001), 'Repeating and remembering foreign language words: Implications for language teaching system', *Artificial Intelligence Review*, vol 16, no 1 , pp 37 - 47.

Diehl, R.L., Lotto, A.J., & Holt, L.L. (2004). Speech perception. *Annual Review of Psychology*, 55, 149-179.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, 49, 296–314.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3–28.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13, 361-377.

Liberman, A. M., & Mattingly, I. G. (1985). "The motor theory of speech perception revised". *Cognition* 21 (1): 1–36.

Luce RD. (1986). *Response times: Their role in inferring elementary mental organization.* Oxford University Press; New York.

Porter, R., & Castellanos, F. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America,* 67, 1349–1356.

Reisberg, D, McLean, J. & Goldfield, A. (1987). Easy to Hear to Understand: A Lip-Reading Advantage with Intact Auditory Stimuli. *In Dodd, B. and Cambell, R. (eds.) Hearing by Eye: The Psychology of Lip-Reading*, 97-113. London: Lawrence Erlbaum.

Schwartz, J.L., Basirat, A., Ménard, L., & Sato, M. (2010). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics,* 25, 336-354.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*. 17 : 2387-2399.

Stevens, K.N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. *In: Wathem-Dunn, W. (ed), Models for the Perception of Speech and Visual Form. Cambridge, MA: MIT Press.*

Sumby W.H., & Pollack I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26 (2), 212-215.